딥러닝 기반 특허의 종속 청구항 인식 개선

박주연*, 신예지*, 김민수*, 김동호**, 김지희**
*동국대학교 컴퓨터공학과
**동국대학교 융합교육원
pjuyeon25@gmail.com, gjsld1@naver.com, duqrlpig@gmail.com
dongho.kim@dgu.edu, jihie.kim@dgu.edu

Improving Recognition of Patent's Claims with Deep Neural Networks

Ju-yeon Park*, Yeji Shin*, Minsu Kim*, Dongho Kim**, Jihie Kim**
*Dept. of Computer Science and Engineering, Dongguk University
** Dongguk Institute of Convergence Education

Abstract

특허를 통해 기술의 권리를 정의하고 보호하는 일이 매우 중요해짐에 따라 특허 문서를 분석하는 연구 또한 중요해지고 있다. 특히 특허의 청구항을 종속항과 독립항을 구분하고, 관련된 인용을 찾아내는 일은 관련 특허들을 분석하는데 매우 중요하다. 본 연구는 최근 텍스트 분석 분야에 획기적 성능 개선을 이끈 BERT(Bidirectional Encoder Representations From Transformers) 언어 모델을 사용하고 Neural Network 의 파인 튜닝 과정을 통해 청구항의 독립과 종속을 구분하였고, 인용하는 항의 번호와 인용 문구로 이루어진 인용 패턴을 통해 종속항의 인용 항을 찾아내었다. 이 방법을 2003 년 이후의 xml 형식의 미국 특허 데이터에 사용한 결과, 정확도 99% 의 성능을 확보하였다.

1. Introduction

4 차 산업혁명 시대를 맞이하여 기술은 점점 빠르게 변화하고 발전하고 있다. 이러한 시대의 중심에서 기술의 권리를 정의하고 보호하는 일은 매우 중요하다. 최근 글로벌 기업들 간의 대규모 특허 침해 소송이 증가함에 따라 특허문서분석을 활용한 체계적인 연구도 요구되고 있다. 특허 권리 범위의 정확한 해석을 위해서는 권리범위에 대한 상위/하위 포함 관계의 명확한 구분이 필요하다. 청구 범위는 독립항과 종속항들로 구성되는데, 이러한 독립항과 종속항들의 명확한 구분을 통해 특허문서분석에 도움을 줄 수 있다.

최근의 특허는 XML/SGML 형태로 제공되어, 인용된 종속항이 태그로 표현된다. 미국의 특허 데이터 관련, 과거의 2003 년 이전의 특허는 텍스트 정보로, 인용하는 항의 태그가 되어있지 않기 때문에 직접 판별해야 한다. 또한, 종속 관계는 다양한 표현으로 이루어져있고, 오기가 있을 수 있기 때문에 주의해야 한다.

따라서 본 연구는 2003 년 이후의 태그 청구범위가 구분된 5 만여개의 데이터를 통해 2003 년 이전 특허 데이터의 청구항 인용 관계를 명확하게 하는 것을 목표로 한다. 특허의 청구항의 상하 포함 관계 파악은 총 2 단계로 이루어지며, 약 8 : 1 의 종속항과 독립항 간의 학습데이터의 양과 질의 균형을 향상시키기 위해 최근 텍스트 분석 분야에 획기적 성능 개선을 이끈 BERT(Bidirectional Encoder Representations From Transformers)[1] 모델을 사용하고, 특허 분류 데이터를 이용하여 파인 튜닝 한다. 문맥의 구조뿐만 아니라 문맥의 이해를 포함시킨 BERT 모델을 통해 특허 청구항의 종속 관계를 명확하게 구분하고, 데이터 불균형 문제도 완화시킨다. 마지막으로 인용하는 항의 번호와 인용 문구로 이루어진 인용 패턴을 이용하여 종속항의 인용 항의 번호를 찾는다. 이 방법을 2003 년 이후의 xml 형식의 미국 특허 데이터에 사용한 결과, 정확도 99% 의 성능을 확보하였다

2. Related Work

특허를 분석하는 다양한 연구들이 진행되고 있는 가운데, 이 연구에서는 특허의 청구항들을 분석하여 독립항과 종속항을 구분하고자 한다.

이미 학습된 BERT 모델의 파인 튜닝을 이용하여 특허의 분류를 진행하는 연구도 있다.[2] 이 연구에서는 특허의 다른 부분이 아닌 청구항 만을 분석하여 특허의 분류를 진행한다. 같은 BERT 모델의 파인 튜

닝을 이용하여 청구항을 분석하지만 최종 분류하는 결과물이 특허 전체의 분류라는 점에서 차이가 있다.

청구항의 분류 알고리즘을 통해 특허 청구항의 구조분석을 하는 연구도 존재한다.[3] 이 연구에서는 청구항의 형식적 특징과 청구대상(subject-matter)을 이용한 독립항과 종속항을 분류하는 알고리즘을 제안한다. 하지만, 이 연구에서는 국내 특허 문서에 한정되며, 청구항의 "n 항(청구항 n)에 있어서/에서"의 문구의 여부를 통해 단순하게 종속항과 독립항을 구분하게 된다. 우리의 연구에서는 BERT 모델을 도입해 보다 종속항을 구분 한다는 점에서 좀 더 높은 완전성을 보여줄 수 있다.

딥 러닝 을 활용한 특허를 분류하는 연구[4]에서는 특허 문서의 IPC(International Patent Classification) 와 IPC sub 분류를 진행한다. 특허 문서 분석을 위해 특징 벡터들을 추출한 후, 인코더 층과 네트워크 층을 활용하여 특징 학습을 진행한 후에 Softmax 회귀를 통해 분류를 하게 된다. 이 연구에서는 특허 문서 분석에 쓰이는 Softmax 회귀 이외에도 특허 문서의 특징을 찾는 학습 과정이 필요하게 되므로 특허 문서 분석 이전의 전처리 과정이 복잡하게 된다.

문장의 형태소 단위의 분석을 통해 특허 항간의 구조 분석을 진행한 연구도 있다.[5] 일본어 특허를 기반으로 문장의 단어에 따라 총 6 개의 문장 구조로 구분하고, 이런 구조를 바탕으로 청구항을 트리 형태로 시각화하여 특허의 가독성을 높인다. 따라서, 특허 청구항에 대한 더 정확한 이해를 할 수 있지만, 이 연구에서는 단순히 청구항의 구조만을 나타낼 뿐 어떠한 인용 관계도 알 수 없기 때문에 청구항의 전체적인 맥락을 이해하는 데에는 어려움이 있다.

이 논문에서는 특허의 청구항을 분석하여 인용 문구와 인용하는 항의 번호로 이루어진 인용 패턴을 이용하여 BERT 학습 모델에 적용시킨 후, 이 학습 모델을 이용하여 간단하게 청구항의 인용 관계를 파악하여 특허 청구항의 내용을 보다 정확하게 파악하는 데 도움이 되고자 한다.

3. Approach

현재 미국의 특허는 청구항에서 인용하는 항의 정보를 태그 형식으로 (그림 1)과 같이 제시하고 있다.

<PATN> US06844315B1
<RN> US6844315
<CLMS> <CLAIMS><CLAIM ORDER='1' id='CLM-00001'> <claim-text>1. A method of treating sepsis in a subject comprising: <claim-text>administering, in a pharmaceutically acceptable manner, a pharmaceutically effective amount of two or more immunoregulators, immunoregulators peptides, functional fragments or functional analogues thereof to the subject, wherein said immunoregulators comprise peptide and recombinant hCG. </claim-text></claim-text> </CLAIM>
<CLAIM ORDER='2' id='CLM-00002'> <claim-text>2. The method of <claim-ref idref="CLM-00001">claim 1</claim-ref> wherein the peptide is selected from the group consisting of SEQ ID NO: 1, SEQ ID NO: 2, SEQ ID NO: 3 and a functional fragment thereof.</claim-text> </CLAIM> <CLAIM ORDER='3' id='CLM-00003'>

(그림 1) 미국 특허 문서 예시

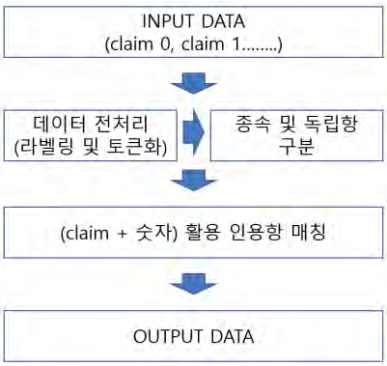
항의 인용 관계 정보를 나타낼 수 있는 태그는, 항의 정보를 가지고 있는 <CLAIM ORDER>와 인용 관계를 제시하는 <claim-ref>태그로, 이 인용 태그들을 중심으로 분석하였다.

인용 태그와 인용 태그의 앞, 뒤 문장들을 (그림 2)와 같이 파싱 하였다. 데이터를 분석한 결과, 각 data 종속항 중 인용 태그이지만 [claim + 숫자] 문구가 없는 경우는 총 1,618,116 개의 데이터 중 7696 개로 약 0.005%에 불과하였다. 따라서, 독립항과 종속항만 잘 구분해낸다면, 산술적으로 99.5% 이상의 확률로 정확한 인용 태그 지침 가능할 것이라고 예상하였다.

PDAT1	Attribute:CLREF	PDAT2
3. Indenyl compound according to	CLM-00001	, wherein R contains at least one aryl group.
4. Indenyl compound according to	CLM-00001	wherein R contains at least one phenylene group.
5. Indenyl compound according to	CLM-00001	wherein R contains a bisaryl group.
6. Indenyl compound according to	CLM-00001	wherein R is a 2,2'-biphenylene.
7. Indenyl compound according to	CLM-00001	wherein M is Ti, Zr or Hf.
8. Indenyl compound according to	CLM-00001	wherein Q is Cl or a methyl group.

(그림 2) 데이터 파싱 예시

이와 같은 [claim + 숫자] 특징을 활용하여 (그림 3)과 같이 청구항 분류의 전반적인 프로세스를 구축하였다.



(그림 3) 전반적인 프로세스

각 종속항과 독립항을 구분하기 위해서 딥러닝 모델을 활용하고자 하였으며 우선 CNN 모델을 활용하였다. 각 청구항의 네번째 이상의 어절을 더미 데이터로 보고 잘라내어 종속 여부를 라벨링 한 뒤 Lookup 테이블 (벡터화 된 입력 데이터)을 구축해 전

처리과정을 마쳤다. 하지만 CNN 모델의 트레이닝 결과는 예상보다 낮은 70% 수준으로 나타났으며, 이에 대한 원인으로서 문장 구조 기반의 알고리즘과 데이터의 편향성(true: 90%, false: 10%)을 지목하였다.

첫번째 문제점인 문장 구조기반의 알고리즘을 개선하기 위해 CNN 모델에서 문맥이해 기반의 알고리즘인 NLP 기반 딥러닝 모델로 변경하였다. 또한 데이터 편향성을 개선하기 위하여 사전 학습된 모델을 튜닝하는 파인 튜닝 방식을 활용하였다. 이때 활용한 모델은 BERT 모델이며 BERT 는 구글이 공개한 사전 훈련된 자연어 처리의 딥러닝 모델이며 일부 성능 평가에서 인간보다 높은 정확도를 보인다. 이러한 사전 훈련된 모델을 튜닝하여 활용한다면 데이터의 부족, 비용 및 시간적인 문제들을 다수 해결할 수 있다.

BERT 모델을 튜닝하기 위해선 해당 모델이 이해할 수 있는 형태로 데이터를 전처리 하는 과정이 필요하다. 우선 모델 학습을 위해 활용될 학습 데이터 셋(약 100,000 여개)을 .tsv 형태의 확장자로 준비하였다. 다음으로 해당 데이터는 (그림 3)과 같이 column 0 에 각 행의 ID, column 1 에 각 행의 라벨(종속성 여부), column 2 에 alpha 데이터, column 3 에 각 행의 문자열의 형태로 구성하였다.

	id	label	alpha	text
0	0	0	a	Unfortunately, the frustration of being Dr. Go...
1	1	1	a	Been going to Dr. Goldberg for over 10 years. ...
2	2	0	a	I don't know what Dr. Goldberg was like before...
3	3	0	a	I'm writing this review to give you a heads up...
4	4	1	a	All the food is great here. But the best thing...

(그림 3) 전처리 데이터 구성

현재까지의 전처리 된 데이터를 인간이 이해할 수 있는 문자표현의 형태라고 한다면, 다음으로 BERT 모델이 이해할 수 있는 특징표현의 형태로 추가 전처리 과정을 진행하였다. 각 행은 라벨과 텍스트의 토큰화가 진행되며 각각의 토큰 쌍은 앞선 column 0 의 id 에 따라 구분된다.

```
{
  "attention_probs_dropout_prob": 0.1,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "max_position_embeddings": 512,
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "type_vocab_size": 2,
  "vocab_size": 28996
}
```

(그림 4) Training parameter

앞서 만들어진 전처리 데이터를 사전 학습 된 BERT 모델에 적용하여 retraining 과정을 거쳐 종속항과 독립항을 구분하는 모델을 만들어내었다. 해당 모델의 테스트 과정을 위해 약 23,000 여개의 테스트 데이터 셋을 준비하였고 앞선 전처리 과정을 거친 후에 evaluate 하였다.

```
# Load pre-trained model (weights)
model = BertForSequenceClassification.from_pretrained(BERT_MODEL, cache_dir=CACHE_DIR, num_labels=num_labels)
# model = BertForSequenceClassification.from_pretrained(CACHE_DIR + 'cased_base_bert_pytorch.tar.gz', cache_dir=CACHE_DIR, num_labels=num_labels)
```

```
  1%|▌
| 3306496/404400730 [00:19<08:08, 820603.15B/s]
```

(그림 5) BERT 모델의 retraining

4. Result

NLP BERT 모델의 튜닝 결과는 다음과 같다. 10 만개의 데이터를 학습시킨 후, 2 만개의 데이터를 통해 테스트를 진행한 결과는 <표 1>,<표 2>와 같다.

항목	값
MCC (Matthews correlation coefficient)	99.97%
Evaluate loss	0.07%

<표 1> BERT 모델 성능

		실제 결과 (종속항 19242 개, 독립항 4351 개)	
		true	false
분류 결과	true	19240 개	0 개
	false	2 개	4351 개

<표 2> BERT 모델 테스트 결과

<표 1> 높은 매튜 상관계수(MCC) 값으로 테스트 데이터가 분류되었다. <표 2> 분류 결과에서는 총 19242 개의 종속항 중 실제로 종속으로 구분한 개수는 19240 개, loss 는 2 개가 나왔다. 또한, 총 4351 개의 독립항 중 실제로 독립으로 구분한 개수는 4351 개, loss 는 0 개가 나왔다. 결론적으로 0.07%의 낮은 손실률(Evaluate loss)로 테스트 데이터가 올바르게 분류되었다는 것을 알 수 있다.

5. Discussion

특허의 청구항들을 독립항과 종속항으로 구분하고, 종속항의 경우 인용하고 있는 항을 찾는 BERT 학습 모델의 정확도가 99% 이상으로 높게 나왔다. 특허의 종속항들이 대부분 인용 문구와 인용하는 항의 번호 형식으로 인용 패턴이 정형화되어 있기 때문에 단순한 BERT 학습 모델을 통해서도 높은 정확도를 얻을 수 있었다. 그렇지만 인용하는 항의 번호가 명시되어 있지 않거나, 인용하는 번호가 여러 개(e.g. 1,2,3 or 1-3)인 예외적인 경우가 존재하는데, 지금의 BERT 학습 모델로는 이런 예외 케이스들을 정확하게 찾을 수 없다. 따라서 정형화되어 있지 않은 소수의 예외 케이스들을 처리할 수 있도록 하기 위해 단순한 패턴을 통해 인용 항을 찾는 것이 아니라 문맥의 이해를 포함한 자연어 처리를 활용한 청구항 인용 방식을 학습하여, 이를 찾아낸다면 하나 이상의 청구항들을 모두 찾아낼 수 있게 되어 더 높은 정확도를 얻을 수 있을 것이다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음" (2016-0-00017)

참고문헌

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 24 May, 2019, Google AI Language, 10 Apr, 2020, <https://arxiv.org/abs/1810.04805>

[2] Jieh-Sheng Lee, Jieh Hsiang, “Patent BERT: Patent Classification with Fine-Tuning a pre-trained BERT Model”, 1 Jul, 2019, National Taiwan University, 10 Apr, 2020, <https://arxiv.org/abs/1906.02124>

[3] 송민호, 임소라, 권용진 “특허 청구항의 구조분석을 위한 청구항 분류 알고리즘”, 한국통신대회, 2018, 102-103(2 pages)

[4] Bing Xia, Baoan LI, Xueqiang LV “Research on Patent Classification Based on Deep Learning”, Advances in Intelligent Systems Research, 2016

[5] Akihiro S., Manabu O., Yuzo M., Makoto I. “Patent Claim Processing for Readability”, “03: Proceedings of the ACL-2003 workshop on Patent corpus processing”, 2003, 56-65(10 pages)